# A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case

Omar Hasan[1], Benjamin Habegger[1], Lionel Brunie[1], Nadia Bennani[1], Ernesto Damiani[2]
[1] University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
{omar.hasan, benjamin.habegger, lionel.brunie, nadia.bennani}@insa-lyon.fr
[2] Department of Computer Technology, University of Milan, Italy
ernesto.damiani@unimi.it

***Abstract*—User profiling is the process of collecting information about a user in order to construct their profile. The information in a user profile may include various attributes of a user such as geographical location, academic and professional background, membership in groups, interests, preferences, opinions, etc. Big data techniques enable collecting accurate and rich information for user profiles, in particular due to their ability to process unstructured as well as structured information in high volumes from multiple sources. Accurate and rich user profiles are important for applications such as recommender systems, which try to predict elements that a user has not yet considered but may find useful. The information contained in user profiles is personal and thus there are privacy issues related to user profiling. In this position paper, we discuss user profiling with big data techniques and the associated privacy challenges. We also discuss the ongoing EU-funded EEXCESS project as a concrete example of constructing user profiles with big data techniques and the approaches being considered for preserving user privacy.**

***Keywords*—User profiling, recommender systems, big data, privacy, EEXCESS.**

## I. Introduction

A user profile is a collection of information that describes the various attributes of a user. These attributes may include geographical location, academic and professional background, membership in groups, interests, preferences, opinions, etc. User profiling is the process of collecting information about a user in order to construct their user profile.

User profiles are utilized by a variety of web based services for different purposes. One of the primary uses of user profiles is for recommendation of items, elements or general information that a user has not yet considered but may find useful. General purpose social networks such as Facebook.com use a user profile to find potential friends based on the existing relationships and group memberships of the user. Professional social networks such as LinkedIn.com exploit the skills and professional background information available in a user profile to recommend potential employees. Search engines such as Google.com use the history of user searches to personalize the current searches of the user.

Big data techniques are a collection of various techniques that can be used to discover knowledge in high volume, highly dynamic, and highly heterogeneous data. Big data techniques offer opportunities for user profiling that can result in very comprehensive user profiles. Big data techniques have two strengths in particular that enable collecting accurate and rich information for user profiles: (1) Big data techniques process unstructured data as well as structured data. Unstructured data of different varieties generated by users is growing in volume with high velocity and contains lots of useful information about the users. (2) Big data techniques can process high volume data from multiple sources. This enables linking user data from different sources and aggregating them into a single user profile. Moreover, user information from different sources can be correlated to validate or invalidate the information discovered from one source.

On one hand, user profiling with big data techniques is advantageous for providing better services as we have discussed above. On the other hand, user profiling raises a significant threat to user privacy. One can assume that an ethical and trustworthy service would use the information collected in a user profile with the user's explicit consent and only for the benefit of the user. However, services that are less inclined toward protecting user privacy, may use user profiles for a number of purposes which may not be approved by the user and which may result in disclosure of personal information. One example is the utilization of user profile data for targeted advertising [1]. Another example is the selling of personal information in user profiles to third parties for profit. The third parties may then use this private information for commercial or even malicious purposes [2]. Privacy breaches may occur even when a service is willing to protect a user's privacy [3].

The ongoing EU-funded EEXCESS (eexcess.eu) project aims to improve user recommendations by making intensive use of user profiling and therefore collecting detailed information about users. The EEXCESS project has to address various privacy challenges which appear mainly due to the use of big data and related technologies. One of the major challenges is that the EEXCESS architecture is based on a federated recommender system in which future partners may join. The trustworthiness and the intent of these partners are not necessarily known. The information collected and disclosed to recommenders may not, in itself, be sensitive, however, cross-referencing it with external big data sources and analyzing it through big data techniques may create breaches in user privacy. Since, untrustworthy partners may have access to such big data sources and techniques, privacy becomes a clear challenge.

In this position paper, we highlight some of the private content contained in user profiles, the big data techniques that can be used to construct user profiles, and the ongoing research work toward addressing the associated privacy challenges. In particular, we consider the EEXCESS project as a use case.

978-0-7695-5006-0/13 $26.00 © 2013 IEEE
DOI 10.1109/BigData.Congress.2013.13

We present the proposed EEXCESS architecture, the privacy goals, and the approaches being considered to achieve those goals.

## II. User Profile Contents, Big Data Techniques, and Preservation of Privacy

### A. User Profile Contents

The information contained in a user profile can be provided explicitly by the user or alternatively it can be either inferred or mined by the service that manages the profile. Gathering accurate, precise, and rich information is clearly the objective when building a user profile. More and accurate information about a user can indeed help services provide better recommendations.

The most common contents of a user profile include: user interests; user knowledge, user background and skills; user goals; user behavior; user individual characteristics; and user context [4]. We briefly summarize these attributes below. An extended description can be found in [4].

We invite the reader to note while going through the descriptions below that each of these attributes can be considered as a user's private information. In Section II-C, we will further discuss the privacy issues in the context of user profiling.

***Interests.*** The information that can be recorded under this attribute includes a user's professional interests, his interests in hobbies, his interests in entertainment such as music, cinema, books, etc., his interests in sports, as well as his interests in commercial products. If the interests of a user are known, a recommender system can use this information to recommend items that are of the highest interest to the user.

***Knowledge, background and skills.*** This attribute can be used to quantify the knowledge of the user in a given domain. For example, the knowledge that a student has acquired by taking an online course could be measured and recorded. Moreover, professional expertise and skills can also be rated. This information can be used to discover experts in a given domain or conversely to rule out individuals whose knowledge, background, or skills do not correspond to a particular task.

***Goals.*** The goals and intentions of a user represent what he wishes to achieve in a given context. Goals can be classified as short term and long term. For example, a short term goal of a student could be to obtain a high grade in a class whereas a long term goal could be to graduate from college. A recommender system can try to predict the needs of a user given his short term and long term goals and intentions.

***Behavior.*** Users often have repetitive behaviors that can be observed and stored in their user profiles. For example, a user may order pizza online most Tuesdays and purchase an online movie most Fridays. Given this information, a recommender system could suggest pizza deals to the user on Tuesdays and new movies on Fridays. The history of user actions may also be considered under this attribute.

***Individual characteristics.*** The individual characteristics of a user that may be made part of their user profile include personal information such as age, gender, relationship status, address, etc. Knowledge of demographic information is useful information for a recommender system. For example, attributes such as age, gender, and address can have a strong impact on the movies that a person views and likes and thus on the recommendations that should be made.

***Context.*** The different types of contexts include environmental contexts, personal contexts, social contexts, and spatio-temporal contexts. Entities that are located in the vicinity of the user form his environmental context, e.g., things, services, temperature, light, humidity, noise, and persons. Personal context comprises of physiological contexts, such as weight, pulse, blood pressure, hair color, etc., as well as mental contexts, such as mood, stress level, etc. Social context can comprise of information such as friends, neighbors, co-workers, and relatives. Spatio-temporal information is a combination of time, location, and the direction of movement.

We observe again that each of the attributes contained in a user profile described above can be considered as private information. For example, a person may not wish to share information regarding his whereabouts at all times with everyone. Similarly, it may be detrimental for a person to reveal her interests, goals, behavior etc. and thus she may not wish to divulge this information.

### B. Big Data Techniques for User Profiling

We list below some of the big data techniques that can be used for collecting information about a user and building a user profile. An extended list of big data techniques that can be used for user profiling can be found in [5].

It can be noted that many big data techniques are in fact adapted from artificial intelligence and graph theory. However, they take into consideration the added constraints implied by big data: the massive amount of data that often requires distribution over multiple servers or clusters, and the diversity of such data. Many existing big data implementations are algorithms adapted for distributed computation platforms such as Hadoop (hadoop.apache.org).

***Network analysis.*** Network analysis algorithms are used to discover relationships between the nodes in a graph or a network. Network analysis is particularly useful in the context of social networks where important information about the user such as his friends, co-workers, relatives, etc. can be discovered. Social network analysis can also reveal central users in the network, i.e., users who exert the most influence over other users. This information can be used to populate the attributes of social and environmental contexts, individual characteristics, etc. in a user profile.

***Sentiment analysis.*** Sentiment analysis is a natural language processing technique that aims to determine the opinion and subjectivity of reviewers. The Internet is replete with reviews, comments and ratings due to the growing popularity of web sites such as Amazon.com, Ebay.com, and Epinion.com where users provide their opinion on others users and items. Moreover, micro-blogging sites

such as Twitter.com and social network sites such as Facebook.com also hold a large amount of user opinions. The goal of sentiment analysis is to classify user opinions. This classification may be a simple polarity classification, i.e., negative or positive, or a more complex one, e.g., multiple ratings. Sentiment analysis can be used to process unstructured text written by a user to discover their interests, opinions, preferences, etc. to be included into their profile.

***Trust and reputation management.*** Trust and reputation management is a set of algorithms and protocols for determining the trustworthiness of a previously unknown user in the context of his reliability in performing some action. For example, a reputation management system could be used for computing the trustworthiness of an online vendor who may or may not deliver the promised product once he receives payment. The reputation of a user is computed as an aggregate of the feedback provided by other users in the system. Trust and reputation information can be an important part of a user profile. It can convey the user's trust in other users as well as his own reputation in various contexts. This information can be subsequently used as a basis for recommending trustworthy users and avoiding those who are untrustworthy. Trust and reputation management systems can function in conjunction with sentiment analysis for obtaining user opinions and then computing trustworthiness and reputation.

***Machine learning.*** Machine learning is a sub-field of artificial intelligence that aims to build algorithms that can make decisions not based on explicit programming but instead based on historical empirical data. An example often cited is the algorithmic classification of email into spam and non-spam messages without user intervention. In the context of user profiling, machine learning can be used for learning user behavior by identifying patterns. Topics in machine learning include: supervised learning approaches, e.g., neural networks, parametric/non-parametric algorithms, support vector machines, etc.; and unsupervised learning approaches, e.g., cluster analysis, reduction of dimensionality, etc.

***Cluster analysis.*** Cluster analysis is the process of classifying users (or any other objects) into smaller subgroups called clusters given a large single set of users. The clusters are formed based on the similarity of the users in that cluster in some aspect. Cluster analysis can be applied for discovering communities, learning membership of users in groups, etc. Cluster analysis can be considered as a sub-topic of machine learning.

### C. Preservation of Privacy in Big Data Techniques

Big data techniques offer excellent opportunities for more accurate and richer user profiling. However, privacy is an issue that can hinder acceptance by users of user profiling with big data techniques. Therefore, there is a need to develop big data techniques that can collect information for user profiles while respecting the privacy of the users. Such privacy preserving big data techniques for user profiling would raise the confidence of users toward collection of their personal information.

There is a significant amount of research currently in progress to achieve the goal of preserving user privacy while collecting personal information. As an example, we cite the field of privacy preserving reputation management. A privacy preserving reputation management system operates such that the opinions used to compute a reputation score remain private and only the reputation score is made public. This approach allows users to give frank opinions about other users without the fear of rendering their opinions public or the fear of retaliation from the target user. Privacy preserving reputation management systems for distributed environments have been investigated since long [6], however, they pose scalability problems as they require large-scale handling of rapidly changing pseudonyms.

Privacy preserving reputation management systems for centralized environments include those by Kerschbaum [7] and by Bethencourt et al. [8]. The system by Kerschbaum introduces the requirement of authorizability, which implies that only the users who have had a transaction with a ratee are allowed to rate him even though rating is done anonymously. Bethencourt's system lets a user verify that the reputation of a target user is composed of feedback provided by distinct feedback providers (implying no collusion) even when users are anonymous. Hasan et al. [9], [10] propose privacy preserving reputation management systems for environments where the existence of centralized entities and trusted third parties cannot be assumed. Current privacy preserving reputation management systems still face a number of open issues. These include attacks such as self-promotion and slandering, in which a user either submits unjustified good opinions about himself or unwarranted bad opinions about a competitor.

Differential privacy, introduced by Dwork et al. [11], is a recent approach to preserving privacy that has received significant attention. It provides a mathematical process for adding randomness to statistical queries with a quantifiable degree of privacy for individuals joining a database. The framework offers guarantees on the risk of joining a statistical database. However, in practice, differential privacy can render some subsets of the randomized data less useful while poorly preserving the privacy of specific individuals. This has been demonstrated for instance in [12]. Thus, privacy preserving techniques still have much to achieve in order to render personal information of users truly private.

Another well-known approach in privacy preservation of published data is *k-anonymity* [13]. It relies on the distinction of quasi-identifiers and sensitive attributes. Quasi-identifiers are the attributes allowing to determine the identity of the individuals referred to by a record (e.g. age, gender, city). The sensitive attributes (e.g. a disease) are those which should not be linkable to the individuals. A set of records $V$ is said to satisfy *k-anonymity* for the set $A_q$ of quasi-identifiers if for every tuple $t \in V$ there exists $k-1$ distinct records $v_i$ ($i \in [1, k-1]$) such that $\forall i \in [1, k-1] \pi_{A_q}(t) = \pi_{A_q}(v_i)$ (where $\pi_A(r)$ denotes the projection of record $r$ on the attribute set $A$).

## III. The EEXCESS Project

EEXCESS (Enhancing Europe's eXchange in Cultural Educational and Scientific resources) (eexcess.eu) is a European Union FP7 research project that commenced in February 2013. The project consortium comprises of INSA Lyon

(insa-lyon.fr), Joanneum Research (joanneum.at), University of Passau (uni-passau.de), Know-Center (know-center.tugraz.at), ZBW (zbw.eu), Bit media (bit.at), Archäologie und Museum Baselland (archaeologie.bl.ch), Collections Trust (collectionstrust.org.uk), Mendeley (mendeley.com), and Wissenmedia (wissenmedia.de). In this section we present the EEXCESS project to illustrate how user profiling can benefit recommender systems particularly with the use of big data techniques. We also discuss the associated privacy issues and the approaches currently being considered in the project for tackling the privacy problem.

The main objective of EEXCESS is promoting the content of existing rich data sources available throughout Europe. While user context is more and more present, the current response of web search engines and recommendation engines to the massive amount of data found on the web has been to order query results based on some form of popularity. It is evident that the introduction of PageRank [14] in search engines has changed the landscape of online searching. However, this has lead to the effect of having large quantities of valuable content remaining simply unaccessed due to low levels of global popularity but at the same time being of high interest for a particular user. This unseen data is sometimes referred to as "long-tail content" in reference to the long-tail of a power-law distribution which in many cases characterizes the distribution of user interest in particular content.

It is this type of long-tail content that some of the EEXCESS partners are providing. This includes precise and rich content such as museum object descriptions, scientific articles, business articles, etc. Currently, this very specific content has trouble finding appropriate visibility, even though they would be invaluable in the appropriate contexts where fine-grained and precise information is seeked for.

The aim of EEXCESS is to push such content made available by its partners to users when appropriate for them. However, this relies on having a precise understanding of a given user's interests and their current context. Different levels of user profiling can help to characterize a user's interests. In EEXCESS, precise user profiles will allow recommending the appropriate content found in multiple data sources.

### A. Architecture

Figure 1 gives a sketch of the currently envisioned architecture for the EEXCESS project from a privacy perspective. From this perspective, EEXCESS is made of four components: (1) A plugin added to the user's client whose role is to collect and transfer the user's context, trigger recommendation requests and render them through rich visualizations, (2) a privacy proxy which collects the user's privacy policy and ensures that it is respected, (3) a usage mining component allowing to identify common usage patterns and enrich user profiles accordingly, and (4) a federated recommender service composed of individual data-sources hosting a specific data collection. The circled numbers on the figure give the information flow when content is being recommended.

As suggested by the presence of a privacy-proxy, one major goal in EEXCESS is to respect its users' privacy. In particular, no information about a user profile data should leak out of the system without the user's consent. As will be

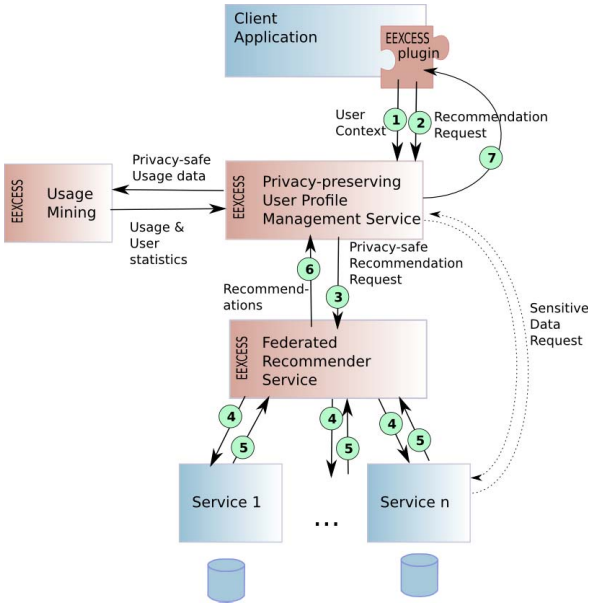


Fig. 1. EEXCESS architecture from a privacy perspective

discussed later, the project is faced with a conflicting situation in which disclosing more information will allow to improve recommendation quality but will also augment the risk if privacy leaks. The exact internals of the privacy proxy are among the works to be completed during the project's time span. For simplicity, we consider the proxy-service as a single peer in this paper.

Let us consider a typical EEXCESS user scenario. Alice is an economist employed by a consulting firm. She is currently working on a business plan for one of her customers on a market which is new to her. As usual she uses her favorite search engine to investigate on the different actors of the market and in particular the potential competitors for her client. Fortunately, EEXCESS is connected to an economic database, and starts pushing to Alice relevant content from this database, which includes detailed descriptions of companies found in the target market of her client and strategic economic data. Alice requires that a high level of privacy is ensured by the system. In fact, she is legally-tied by a non-disclosure policy with her customer. In particular, it should not be learned that Alice's customer is taking a move toward the new market.

### B. User Profiling

One of the major objectives of EEXCESS is providing its users with quality recommendations. To this extent, fine-grained user-profiling will be an important part of the project and will consist of collecting sensitive data about the user. Many of the attributes discussed in section II-A will be collected or enriched using big data techniques described in section II-B.

Of course, the user's *individual characteristics* will be part of his profile. An EEXCESS user's *interests* will either be interactively collected and/or completed using big data techniques implemented particularly by the usage mining service. User actions will be tracked by the EEXCESS plugin allowing

to keep track of a user's *behavior*. Among the partners of EEXCESS, Bit Media is an e-learning platform. In this case, it is clear that the user's learning *goals* and current *knowledge* (e.g. in the form of courses already taken) will be part of the user's profile. In EEXCESS, the user's *context* will consist of information such as his current geo-location, the document or web page (both URL and content) he is working on, his browsing history, the navigation page which lead to the current page, etc.

To capture an even better understanding of the user, different big data techniques will be applied to further enrich his profile. For example, usage mining will try to identify usage trends, as well as information about the user's unexpressed goals and knowledge. On-the-fly analysis of user interests, context, and expectations is also planned. Clustering techniques may be used to identify communities within EEXCESS users. This profiling and better understanding of the user has a unique goal in EEXCESS of providing the user a personalized experience of the system and in particular *personalized recommendations*. Indeed, the content of the EEXCESS partners being very specific (i.e. being in the long-tail of documents when ordered by popularity), having a fine-grained understanding of EEXCESS user's is essential to link the correct users to the correct content.

In our example, the EEXCESS system will have collected significant information about Alice: her interests (economic information), some comprehension of her goal (writing a business plan), her knowledge (expert in economics), her context of work (information about her customer, the target market, the information she has already collected, etc.). Knowing as much as possible about Alice and her customer will allow the EEXCESS system to provide her with adapted recommendations. For example, instead of presenting general-purpose information about the market, the system will propose more detailed technical data which Alice needs and understands.

### C. Privacy

Providing users with quality recommendations is a seemingly conflicting objective with the equally important goal of privacy preservation. Even a small amount of personal information may lead to identifying a user with high probability in the presence of side channel external data [3].

Returning to our example, it would be unacceptable to Alice that any information about herself or her customer leak out of the system. Alice's project may even be so sensitive that even the fact that *someone* (without particularly knowing who) is setting up a business plan on the target market may be an unacceptable leak because it could lead to competitors taking strategic moves. This emphasizes the fact that preserving only anonymity may not be sufficient in some cases.

Therefore, for EEXCESS to be a success, many privacy-related challenges will have to be addressed.

***Providing privacy guarantees.*** At all levels within the system user privacy guarantees must be given. This is most likely one of the hardest tasks. Indeed, as soon as information flows out of a system, sensitive information leaks become a risk. Solutions which may seem trivial, such as anonymization have been shown to be inefficient. A well known example showing that simple anonymization is insufficient to protect privacy is the de-anonymization of the data of the Netflix contest [3]. Furthermore, Dwork [11] has shown that the published results of a statistical database may lead to privacy breaches even for users who are not originally part of the database. These examples show the difficulties which will have to be overcome in order to provide a privacy-safe system. Furthermore, these works show that research on privacy has shifted from totally preventing privacy breaches to minimizing privacy risks. One of the difficulties to overcome in the EEXCESS project, is to ensure that the collection of information flowing out of the system to potentially malicious peers, limits the risks in breaching any of the users' policies. It goes without saying that the attackers themselves very likely have access to big data techniques and that this aspect should be taken into account.

***Flexible privacy policies.*** Users are different, in particular with respect to privacy. Some may not have any privacy concerns at all where as others may not want to disclose a single piece of information about themselves. For example, in one hypothesis, our user Alice may simply wish to remain anonymous. In another hypothesis, Alice may not be concerned by her identity being revealed, but wish that some information about her be kept private (e.g. she may wish to keep private that she is affected by a particular disease). One big challenge will be to define a policy model which allows for such flexibility and at the same time allows to ensure the policy is respected. Preventing direct disclosure of information marked private is quite straight forward. However, a real challenge is preventing the disclosure of the same information *indirectly*. Indeed, leaking other non-private information of a user's profile can lead, through inference, to unwanted disclosures.

***Evaluating trust and reputation.*** What user profile information is disclosed, or at which granularity it is disclosed, may depend on the trust (with respect to privacy concerns) that the user and/or the EEXCESS system has in the content provider. Calculating a content provider's reputation and trustworthiness in a privacy preserving manner is thus an issue.

Let us consider the case of a user wishing to remain anonymous to all the recommenders. In this case, the attacker could be one of the content-providers trying to collect information about the user that it receives queries from. The EEXCESS privacy requirements for such a user would include:

***Content anonymity.*** To guarantee privacy, the attacker should not be able to identify the user from the provided data. Therefore, the system should ensure that an attacker cannot deduce from the content of a request who it originated from.

***Request unlinkability.*** If multiple queries can be linked together, even while having content-anonymity for each individual query, the combination of the two could reveal information about the user. Therefore, it should be required that the protocols guarantee that two independent requests originating from the same user are unlinkable.

***Origin unlinkability.*** This should be feasible by anonymizing

the origin of the request but under the condition that the origin is not revealed by the application level protocols. Therefore, we also need to guarantee that the application level protocols are privacy-preserving (i.e. an attacker cannot link a given request to the requesting user).

Respecting these three constraints is an ideal goal which requires limiting the information transmitted in each request. Such limitations have a high impact on the utility of the profile information disclosed. Thus the challenge is more to find a balance between privacy and utility than to ensure complete privacy.

In information systems (such as recommender systems, statistical databases, anonymized datasets), the main goal of privacy preservation is to not reveal sensitive information about a single entity within the underlying data. This has been shown to be a difficult goal [11], [15]. In a survey on privacy in social networks, Zheleva and Getoor [16] describe some of the common approaches for preserving privacy: *differential privacy* and $k$*-anonymity*. In the context of recommender systems using collaborative filtering, an approach is to use big data techniques such as clustering to group users together in order to provide privacy [17], [18], [19] with the theory of $k$-anonymity.

In our particular setting, we are faced with a federated recommender system in which trusted and untrusted peers may exchange information. This requires that both the protocols for exchanging information and the content disclosed are privacy-safe. Furthermore, recommendations may not always be limited to a single recommendation technique among the peers. Each content source may wish to use its own approach. In the context of EEXCESS, few hypotheses can be made on the computational capacities or the background knowledge that an untrusted peer may have access to.

Our work in the EEXCESS project will include developing mechanisms for the definition of flexible user privacy policies, guarantees based on the user privacy policies for non-disclosure of private information, quantification of the risk of disclosing private information, mechanisms for exchange of information based on the reputation and trustworthiness of partners, as well as the definition of the relationship between the amount of information revealed and the quality of recommendations.

## IV. Conclusion

In this paper, we discussed the challenges raised when building systems which require at the same time a deep level of user-profiling and a high level of user privacy. Building and disclosing fine-grained user profiles can be highly effective in providing quality recommendations. This is particularly true when recommending long-tail data. Big data techniques play an important role in making these profiles even more specific. On the other hand, this raises the issue of respecting a given user's privacy. Big data may even increase this risk by providing attackers the means of circumventing privacy-protective actions. We illustrated these issues by introducing the challenges raised by EEXCESS, a concrete project aiming both to provide high quality recommendations and to respect user privacy.


## Acknowledgment

The presented work was developed within the EEXCESS project funded by the EU Seventh Framework Program, grant agreement number 600601.


## References


[1] P. Jessup, "Big data and targeted advertising," http://www.unleashed-technologies.com/blog/2012/06/28/big-data-and-targeted-advertising, June 2012.

[2] J. Yap, "User profiling fears real but paranoia unnecessary," http://www.zdnet.com/user-profiling-fears-real-but-paranoia-unnecessary-2062302030/, September 2011.

[3] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, May 2008, pp. 111–125.

[4] S. Schiaffino and A. Amandi, "Intelligent user profiling," in *Artificial Intelligence An International Perspective*. Springer, 2009, pp. 193–216.

[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," The McKinsey Global Institute, Tech. Rep., May 2011.

[6] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, P. Samarati, and F. Violante, "A reputation-based approach for choosing reliable resources in peer-to-peer networks," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2002, pp. 207–216.

[7] F. Kerschbaum, "A verifiable, centralized, coercion-free reputation system," in *Proc. of the 8th ACM workshop on privacy in the e-society (WPES'09)*, 2009, pp. 61–70.

[8] J. Bethencourt, E. Shi, and D. Song, "Signatures of reputation: Towards trust without identity," in *Proc. of the Intl. Conf. on Financial Cryptography (FC '10)*, 2010, pp. 400–407.

[9] O. Hasan, L. Brunie, E. Bertino, and N. Shang, "A decentralized privacy preserving reputation protocol for the malicious adversarial model," *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, vol. DOI: 10.1109/TIFS.2013.2258914, 2013.

[10] O. Hasan, L. Brunie, and E. Bertino, "Preserving privacy of feedback providers in decentralized reputation systems," *Computers & Security*, vol. 31, no. 7, pp. 816 – 826, October 2012, http://dx.doi.org/10.1016/j.cose.2011.12.003.

[11] C. Dwork, "Differential privacy," in *ICALP (2)*, ser. Lecture Notes in Computer Science, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., vol. 4052. Springer, 2006, pp. 1–12.

[12] R. Sarathy and K. Muralidhar, "Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data," *Transactions on Data Privacy*, vol. 4, no. 1, pp. 1–17, Apr. 2011.

[13] L. Sweeney, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, Oct. 2002.

[14] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[15] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, May 2008, pp. 111–125.

[16] E. Zheleva and L. Getoor, "PRIVACY IN SOCIAL NETWORKS: A SURVEY," C. C. Aggarwal, Ed. Boston, MA: Springer US, 2011.

[17] J. Canny, "Collaborative filtering with privacy," in *Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on*. IEEE, 2002, pp. 45–57.

[18] D. Li, Q. Lv, H. Xia, L. Shang, T. Lu, and N. Gu, "Pistis: A Privacy-Preserving Content Recommender System for Online Social Communities," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1. IEEE, Aug. 2011, pp. 79–86.

[19] A. Boutet, D. Frey, A. Jegou, and A.-m. Kermarrec, "Privacy-Preserving Distributed Collaborative Filtering," INRIA, Rennes, Tech. Rep. February, 2013.

# From Context-Aware to Context-Based: Mobile Just-In-Time Retrieval of Cultural Heritage Objects

Jörg Schlötterer, Christin Seifert, Wolfgang Lutz, and Michael Granitzer

Media Computer Science, University of Passau, Germany
{joerg.schloetterer,christin.seifert,michael.granitzer}@uni-passau.de

**Abstract.** Cultural content providers face the challenge of disseminating their content to the general public. Meanwhile, access to Web resources shifts from desktop to mobile devices and the wide range of contextual sensors of those devices can be used to proactively retrieve and present resources in an unobtrusive manner. This proactive process, also known as just-in-time retrieval, increases the amount of information viewed and hence is a viable way to increase the visibility of cultural content. We provide a contextual model for mobile just-in-time retrieval, discuss the role of sensor information for its contextual dimensions and show the model's applicability with a prototypical implementation. Our proposed approach enriches a user's web experience with cultural content and the developed model can provide guidance for other domains.

## 1 Introduction

Recent initiatives like Europeana[1] spend a huge effort on aggregating digitized museum artifacts of different institutions and providing a unified interface to access those resources. Nevertheless, users still need to be aware of those specialized portals to gain access to the tremendous collection of cultural heritage objects. Our approach is to take the content to the user, instead of taking the user to content. To this extent, we implement a just-in-time retrieval approach in a mobile setting, based on the contextual information collected by the various sensors of nowadays smartphones. These sensors capture a wide spectrum of a user's context and hence provide a great source for retrieving relevant resources and adapting to the user's needs. We align our approach along the following questions: *When* to retrieve and present resources to the user? *What* are the resources the user is interested in and can they be refined by location information of resources or users (*where*)?

Specifically, our contributions are the following: (i) we present a context model for just-in-time retrieval in a mobile environment, (ii) we discuss how to incorporate available sensor information into the defined context dimensions and (iii) demonstrate the applicability of the model with a prototype.

[1] http://www.europeana.eu/

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 805–808, 2015.

## 2 Modeling Context for Mobile Just-In-Time Retrieval

Context is usually deemed an additional dimension for personalization, either in a recommender system [2] or in information retrieval [12]. In contrast, context is the sole basis for providing recommendations in our work, resembling just-in-time retrieval [10]. Hence, we follow the rather broad definition of context as "any description of the world that can be relevant to an application" by Pete Steggles [1]. For the task of retrieving relevant resources in a mobile setting we define three abstract dimensions: (i) *when*, (ii) *what*, and (iii) *where*. The rational behind the three dimensions is to construct a conceptual model of the user's (potential) information need, which then can be encoded into a search query. In the following we outline how information for each dimension can be collected from either primary context, i.e. (raw/physical) sensor data (e.g. temperature), or secondary context, i.e. virtual sensors, gathering information from applications or services (e.g. message contents) or logical sensors, gathering information from physical or virtual sensors, mainly by aggregation (e.g. activities, such as walking) [3,6].

***When*:** A user should be notified about additional resources only when it is appropriate. Interruptibility refers to a state, in which a person can be interrupted in a task without (too) negative consequences. Middleton highlights the necessity for Interface Agents to "detect when and if to interrupt the user" [8]. Noise level, observable directly by physical sensors, has been found to be a strong indicator for non-interruptibility [7]. Besides the noise level, interruptibility can be assessed according to the current situation, obtainable from logical sensors. We classify situations into *trigger* and *blocker* situations, that either initiate the recommendation process or hinder it. A combination of situations can also occur, while mostly a blocker situation will supersede one or more trigger situations.

***What*:** One of the most valuable sources for generating search queries is textual content, which is available from the currently used application, incoming messages, notifications, etc. through virtual sensors. In order to translate the textual content into a query, keywords need to be extracted. A first step to separate stopwords and non-informative terms from those that actually convey information is named entity detection [9]. In this process, special challenges of mobile devices need to be addressed, such as short messages [11] or limited resources [4]. Given a candidate set of entities, they can be further reduced, by selecting e.g. the most salient ones [5], matching them against a user profile, etc. A very simple approach, even performable with a mobile phone's limited computing power is to choose based on frequency, i.e. how often an entity is mentioned in the text. The final set of keywords may be enriched with location information (c.f. *where*) and sent to the retrieval system.

***Where*:** Location information also serves as information source to construct or refine a query. In the simplest scenario, the name of the city, the user is currently situated in, can be used as query term, in order to obtain resources about this city. Moreover, based on the current location, points of interest (POIs) nearby can be obtained, and a POI's label can be used as query term. In addition, locations identified by named entity detection (c.f. *what*) can also be used

for retrieval. Cultural heritage objects can exhibit different types of locational information: the actual location of the object, i.e. the museum in whose collection it is stored, the place, from where it originated, etc. Consequently, mapping the detected locations to the appropriate query or metadata fields poses a challenge.

## 3 Prototype

To demonstrate the applicability of our proposed approach, we implemented a prototype[2] for Android mobile devices, which uses the Europeana API as search backend. Figure 1 provides a general overview of the processing chain implemented in our prototype, which is described in more detail in the following.

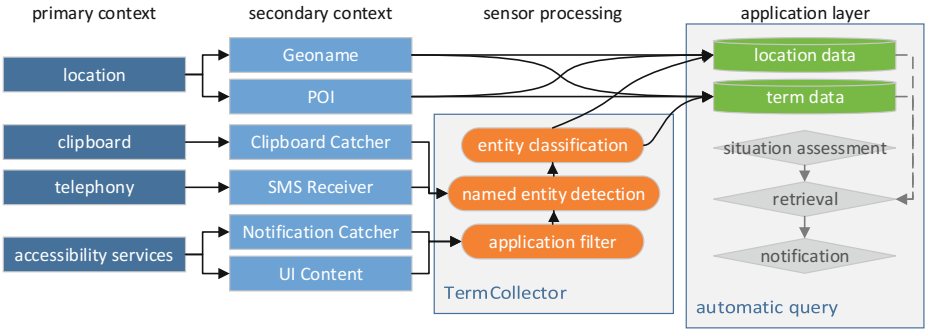


**Fig. 1.** Overview of the prototype's workflow

***When***: We monitor incoming SMS and notifications, the content of the clipboard, content on the user interface (UI) and the user's location. Based on the latter, we obtain the geoname of the current location and nearby POIs. Activity in the just mentioned sensors is a trigger for determining when to query. After processing the context, the automatic query component evaluates the current situation(s) against a predefined set of trigger/blocker situations and, if appropriate, issues a query with the terms derived. If this query yields results, the user is notified through a *ramping interface* [10], featuring different stages, with each stage providing a little more information. The first stages can be ignored easily and information can be filtered early, requiring less attention from a user.

***What & Where***: Context collected from notifications and the UI is filtered first by an application blacklist, in order to remove regular content such as the home screen or notifications from the Android downloader. A simple named entity detection, based on capitalization, is performed for the last four secondary context sensors in the figure and the resulting entities are classified into location or other entities. These steps are not necessary for the POI and Geoname components, as they already provide location entities. The entities obtained from all sensors are stored for further processing by the automatic query component. It

[2] Source and demo at `http://purl.org/eexcess/components/android-app`

is to note, that location entities can also be used to address the *what* dimension as described in section 2. The Europeana API features a faceted search interface, including the facets *what* and *where*. We send the terms stored in the location data component in the *where* facet and those from term data in the *what* facet.

## 4 Summary and Future Work

We presented an approach for mobile just-in-time retrieval in the cultural heritage domain with a retrieval process purely based on contextual information and not requiring any explicit user interaction. We showed how such a process can be modeled along the contextual dimensions of *when*, *what* and *where*, along with a first prototype implementing this model. Even though our application focus is on cultural content, we think that the proposed model can also provide guidance for other domains. In future work, we aim to incorporate the quality of retrieved results into the decision of when to present additional resources to the user instead of relying on a binary decision based on trigger/blocker situations.

**Acknowledgments.** The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

## References

1. Abowd, G.D., Dey, A.K.: Towards a Better Understanding of Context and Context-Awareness. In: Proc. of HUC, pp. 304–307 (1999)
2. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Recommender Systems Handbook, pp. 217–253. Springer (2011)
3. Baldauf, M.: A survey on context-aware systems. International Journal on Ad Hoc and Ubiquitous Computing 2(4) (2007)
4. Ek, T., Kirkegaard, C., Jonsson, H., Nugues, P.: Named entity recognition for short text messages. Procedia-Social and Behavioral Sciences 27, 178–187 (2011)
5. Gamon, M., Yano, T., Song, X., Apacible, J., Pantel, P.: Identifying salient entities in web pages. In: Proc. of CIKM, pp. 2375–2380 (2013)
6. Hong, J.Y., Suh, E.H., Kim, S.J.: Context-aware systems: A literature review and classification. Expert Systems with Applications 36(4), 8509–8522 (2009)
7. Hudson, S.E., Fogarty, J., Atkeson, C.G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J.C., Yang, J.: Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In: Proc. of SIGCHI (2003)
8. Middleton, S.E.: Interface agents: A review of the field. CoRR cs.MA/0203 28 (March 2002)
9. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
10. Rhodes, B.J.: Just-In-Time Information Retrieval. Ph.D. thesis, Massachusetts Institute of Technology (2000)
11. Ritter, A., Clark, S., Mausam, E.O.: Named entity recognition in tweets: An experimental study. In: Proc. of EMNLP, pp. 1524–1534 (2011)
12. Shen, X., Tan, B., Zhai, C.: Context-sensitive information retrieval using implicit feedback. In: Proc. of SIGIR, p. 43 (2005)

# Web-based Just-In-Time Retrieval for Cultural Content

Jörg Schlötterer
University of Passau
Innstrasse 33a
Passau, Germany
joerg.schloetterer@uni-passau.de

Christin Seifert
University of Passau
Innstrasse 33a
Passau, Germany
christin.seifert@uni-passau.de

Michael Granitzer
University of Passau
Innstrasse 33a
Passau, Germany
michael.granitzer@uni-passau.de

## ABSTRACT

Digital content providers of cultural resources face the challenge of disseminating their content to interested users – either because users do not know the existence of resources or do not know the access points. We propose to apply just-in-time retrieval mechanisms to bring the content to the users in a web-based scenario. Our first prototype is a Chrome extension which proposes Europeana content based on the current user context, i.e., the current web site. We think that our approach is promising for scenarios where users are not aware of available content and therefore can and do not explicitly state an information need.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [query formulation, relevance feedback]; H.3.7 [**Digital Libraries**]: [dissemination, user issues]

## Keywords

Cultural Heritage, Personalized Search, Browser Extension, Just-in-Time-Retrieval

## 1. INTRODUCTION

The Web contains a plethora of valuable cultural, scientific and educational resources like scientific papers, tutorials and digitized museum artefacts – mostly hidden in the long-tail of the Web [1]. This means, that a regular web user does either not know the resources exist or cannot find it with a general search engine like Google because of the lower popularity of those sites reflected in the search ranking. Specialized search interfaces like Europeana[1] and Collections Trust[2] aim to bridge this content-user gap and provide access to long-tail contents. While this greatly improves the accessibility of the content, from the users' perspective it still requires the knowledge of the specialized search engine and an explicit formulation of a search query.

A recent initiative, so called edit-a-thons enhance the online encyclopedia Wikipedia with external resources, e.g. cultural content from digital libraries. Their goal is to bridge the content-user gap by bringing together Wikipedians with providers of long-tail contents, such as museums, libraries and archives and disseminating their contents via a major web hub (namely Wikipedia). Europeana, the European aggregator for digital, cultural heritage content, co-organized Wikipedia edit-a-thons in 2012 and 2013. In a case study they summarize the experiences in their 3 edit-a-thons with being successful by enhancing the Wikipedia and providing higher quality content to the public, content providers get a new distribution channel for their content [7].

Current work in personalized access to cultural heritage mainly focuses on server-side [4] or mobile solutions [10]. While the former requires additional effort from content-providers, which may be not feasible for small digital libraries, the latter applies to (mobile) application areas, differing from a typical web browsing setting. We investigate how access to cultural heritage can be realized in a typical web browsing setting without burdening content providers.

In information retrieval research, just-in-time information retrieval encompasses approaches to automatically present search results to users based on the users' context in a non-intrusive manner [9]. In just-in-time information retrieval, the user does not explicitly state her information need, i.e. does not formulate an explicit query. Studies showed that users access more information than with traditional search engines [9].

In previous work [3] we described two general usage scenarios for digital cultural heritage: content-consumption and content-creation. In the content-consumption scenario the user accesses the resources for personal use, whereas in the content-creation scenario the content is integrated or linked in a web page or online social media (blog, news article, online encyclopedia, tweet) and thus disseminated further. In this paper we report work in progress to just-in-time retrieval of cultural content in a web-based setting for the content-consumption scenario. We propose to use a web browser extension which observes the current context of the user (the page) and presents relevant resources – if available.

First, we describe the usage scenario in detail in section 2. The general approach and design decision to solve this scenario are presented in section 3, followed by our prototypical implementation in section 4. We conclude and provide an outlook on future directions in section 5.

[1] `http://europeana.eu`
[2] `http://www.collectionstrust.org.uk/`


*PATCH* '14 Haifa, Israel

## 2. DETAILED SCENARIO

We follow the Scenario-Persona approach, where a Scenario is defined as "a concise description of a persona using a software-based product to achieve a goal" [2]. Our persona is represented by Bob K., a digital native, experienced in using web-based tools and not afraid of using new technologies. Bob has no experience on using the facilities of cultural heritage providers and in general only limited knowledge in the cultural heritage domain.

To acquire information on a certain topic, he will probably start with Google or Wikipedia and has no knowledge about specialized search interfaces like Europeana or web sites of memory organizations in his area. We assume that either way he will end up on Wikipedia to do his inquiry. Most likely, the resources found on Wikipedia might give him a direction of how to satisfy his information need, providing basic information and maybe some historic background-information. But it will not suffice to get a complete picture, he will need more detailed information.

The just-in-time retrieval process for cultural content will utilize the results provided by Wikipedia and maybe also the initial query to form further queries, sending the queries to different content providers and aggregating their results. It furthermore has information on Bob's location via the browser's Geolocation API [8]. The returned resources will be presented to Bob in an appropriate way, according to the nature of the underlying data. These presentations can range from simple textual result lists to graphical visualizations of the results. e.g., a time line.

The search system also returns links to institutions (i.e. museums) that provide content and/or documents on the subject (e.g. photos) or maybe even hold a relevant object at their exhibition. As a result, Bob will not only have information on the subject in general, but he will also be provided with information on (i) specific objects that are related to the subject and held by local museums, (ii) where to look at/use those objects in a museum nearby, (iii) where to get additional, first-hand information on those objects.

While we described a very specific scenario here, in general the retrieval process is not limited to Wikipedia, but will be triggered on every web page for which additional information is desirable and provide the desired information.

## 3. APPROACH

To solve the scenario described above, i.e., to enrich the user's browsing experience with cultural content, additional functionality needs to be added to existing web pages, which is called "JavaScript injection". By executing injected JavaScript, the look and feel of web pages can be altered and additional information and functionality can be added. This injection can be achieved by (i) a bookmarklet, executing JavaScript commands stored as bookmark, (ii) a browser extension - once installed, extending the browser's functionality on all web sites, or (iii) a widget, embedded into a web page server-side.

Although the initial burden to install a browser extension is a little higher compared to the other approaches (while the bookmarklet is simply stored as bookmark, the widget requires no user action at all), we chose to implement a browser extension because of the following advantages: The widget is limited to the web pages implementing it and the bookmarklet needs to be triggered on every page on which additional information is desirable, whereas the browser extension is applicable to all web pages without additional effort. Also, both, bookmarklet and widget are limited to the context of the current page and cannot account for additional information sources (e.g. browsing history), to tailor the injected cultural contents towards the user's interests.

After examining the extension possibilities of the major web browsers and their market share, we decided to start with an extension for Google Chrome for the following reasons: According to StatCounter[3] Google Chrome has the highest (and still growing) market share of 40%. Moreover, Firefox, Safari and Opera share similar extension architectures with Chrome, all based on standard web technologies, such as HTML, JavaScript and CSS. Thus, the results from a Google Chrome extension can easily be transferred to the aforementioned other browsers. Developing an extension for Google Chrome and porting the result to the browsers with a similar extension architecture will cover around 70% of the browsers in use.

## 4. IMPLEMENTATION

Our first prototype of an extension for the Google Chrome browser[4] provides personalized search results from the cultural heritage domain, based on the contents of the current web page or the contents of a selected paragraph within this web page. As back end for the automatically generated queries, we use the Europeana API, which can be easily exchanged by a more sophisticated personalized search system (aggregating information of different content providers and incorporating usage feedback) later on.

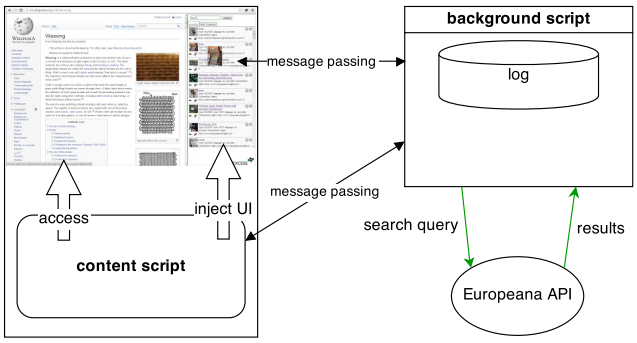


**Figure 1: Architecture of the Google Chrome extension prototype**

### 4.1 Architecture

Figure 1 shows an overview of the architecture. The basic architecture is already predetermined by the extension API of Chrome, partitioning the extension in (i) a *manifest file*, containing information about the extension and configuration settings such as permissions, (ii) *content scripts*, injected into web pages and having access to the particular DOM-tree and (iii) a *background page or script*, which is a single (long-running) instance of the extensions logic with access to the browser API.

The content script is responsible for retrieving and preprocessing the contents of the current page and forwarding

[3] http://gs.statcounter.com/

[4] http://mics.fim.uni-passau.de/demos-downloads/

the results to the background script, as well as for injecting the user interface into the page. The background script creates a query, based on the contents received from the content script, sends this query to the search back end and forwards the results to the injected user interface. The background script furthermore stores the user's interaction and feedback in a log.

## 4.2 User Interface

Figure 2 shows the basic user interface, injected in every web page, simulating the behaviour of a sidebar. The user can toggle the visibility of the interface globally for all pages with the EEXCESS icon 01.The search terms which triggered the presented results are displayed in the search field 02 and can be edited by the user. Adjustments to the query are logged, in order to be able to learn the user's query preferences. Each result item 03 is presented with a preview image 04, the title 05 and facets, such as type, language, provider and rights (if available) 06, along with further interaction possibilities: By clicking either the title or the preview image, an overlay is shown with an HTML representation of the resource, containing additional information. Presenting this information inside the current web page, enables us to keep track of dwell time on a particular result directly within the extension and does not draw the user's attention away from his initial intention.

Beneath storing potential views of a result item as implicit feedback, the user can give explicit feedback, by rating the result up or down with the buttons at 07. These ratings are stored in the Open Annotation format [11] and thus can be easily shared across different platforms. The link button 08 provides a reference URL of the resource for bookmarking or sharing it in an online social network.

The "options" tab 06 is a shortcut to the extension's options page, which is accessible via the extension administration page in Chrome as well. At the current state, it solely provides the ability to enter and edit demographic information.

## 4.3 Content Script

Separate instances of the content script are injected into every web page. Since the content script has full access to the DOM-tree of the particular page, it is responsible for mining the contents of the page and forwarding them to the background script. In the first approach, the three most frequent words of the current page or selected paragraph (if any) are used as query terms to retrieve recommendations.

Beneath mining the contents of the given web page, the content script also monitors the user's interactions with the page, such as mouse clicks or textual input. These interactions can serve as additional information for deducing the query terms. They allow us for example to establish search histories from search engines the user habitually uses. While search histories provide a great information source for personalization strategies [12], we do not expect users to extensively utilize the search interface in our injected sidebar (and do not intend having users formulate a massive amount of queries, but instead providing interesting recommendations automatically). Thus, by keeping track of the user's inputs, we have search histories available without any additional effort of the user.

The third task of the content script is the injection of the user interface (described in section 4.2) into the current web page via an *iframe*-element. The use of an iframe prevents the interface from directly interacting with the current page, due to the cross-origin policy, but is necessary to avoid inheriting CSS-styles of the current page. An iframe-element is advantageous in a further aspect as well: it allows the injection of other user interfaces, which are totally decoupled from the basic one. Since the injected user interfaces communicate directly with the background script via message passing, additional interfaces only need to implement the interface to the background script and no other component.

## 4.4 Background Script

Only a single instance of the background script exists for the whole extension. This script is responsible for the communication with the search back end and serves as a mediator between the user interface injected in a web page and the respective content script, since these two cannot communicate directly. User related information, retrieved from a content script (such as interactions on a web page or aggregated contents of this page) gets logged by the background script. Based on these logs, we plan to establish user profiles, in order to improve search result personalization. The logs contain also usage feedback about recommended resources, such as the viewing time of a recommended result or a rating for it.

To overcome limitations of the browsing history in the browser's API, the background script features an own history implementation, containing not only timestamps for the beginning of a visit on a certain web page, but instead storing the active dwell time on that particular page. "Active" in this case means, the browser and the tab with the particular page within the browser has the focus (switching to another application means switching the focus and thus, the visit ends). In addition, the referring URL (if any), important for deducing navigational paths or patterns, is stored explicitly along with the visit, while retrieving it via the Chrome API is quite cumbersome.

## 4.5 Data Storage

At the current point of time, all data handled within the prototype is stored locally on the client, since this is advantageous in terms of privacy concerns: the user has full control over the data stored and all the information resides within the client, not being transferred to any outside system. The drawback of this approach is that client-side stored logs are prone to be deleted by accident: when the user deletes her private data via the browser integrated function, all of the extension's data are swept away as well.

Basically, two mechanisms exist for storing data on the client: via Web Storage[5], depositing the data as stringified key-value pairs, or in an Indexed Database[6], which consists of object stores, holding records of key-value pairs. We decided to go with the latter, as it provides some significant advantages over Web Storage: In terms of the keys, additional data types, such as Numbers, Date- and Array-objects are allowed. Also, duplicate values for keys can be stored (and iterated). The values to be stored must be supported by the structured clone algorithm, providing some benefits over JSON-serialization, such as being able to handle Blob-, File- and FileList-objects for example. The main advantage of the Indexed Database over Web Storage is the efficient retrieval of records via indexes.

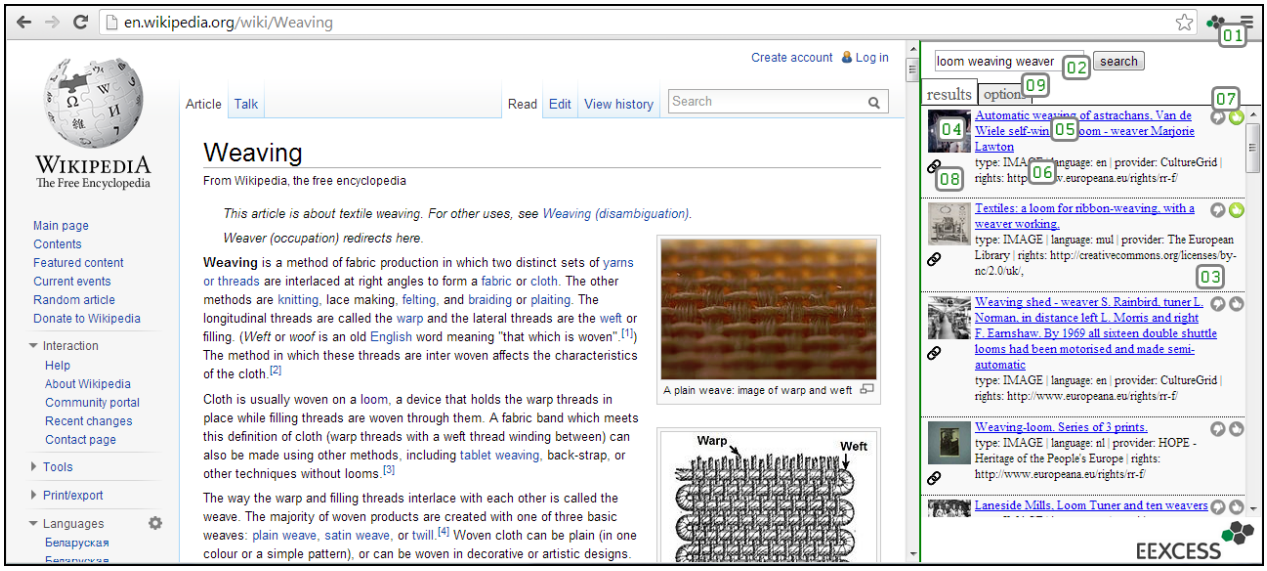


**Figure 2: Screenshot of injected user interface**

## 5. CONCLUSIONS AND FUTURE WORK

We described challenges and opportunities for enriching a user's browsing experience with cultural contents in a web setting and presented our approach to it: a first prototype of a Google Chrome extension for just-in-time retrieval. In the future we will include more sophisticated context detection and query (re-)formulation methods. We will establish user profiles, capturing a user's interests and identify the need for additional information by means of task detection. Further, we will research alternative result presentation methods, e.g., using information visualizations providing an overview of the retrieval results. In addition, we aim to develop a standardized data model and API for connecting additional data providers. We plan to evaluate the usability of the interface with a group of end users and the quality of recommendations with domain experts.

Following the current debate on privacy aspects of user data we will provide means for users to adapt their privacy settings to their personal need within the extension. Further, we will research privacy-preserving personalization approaches, e.g., by not personalizing for single users but for user groups and thus guarantee k-anonymity.

## 6. ACKNOWLEDGEMENTS

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

## 7. REFERENCES

[1] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1–4):69 – 77, 2000.

[2] A. Cooper. *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity*. Pearson Higher Education, 2004.

[3] M. Granitzer, C. Seifert, S. Russegger, and K. Tochtermann. Unfolding cultural, educational and scientific long-tail content in the web. In S. Berkovsky, E. Herder, P. Lops, and O. C. Santos, editors, *UMAP Extended Proceedings*, volume 997, 6 2013.

[4] C. Hampson, E. Bailey, G. Munnelly, S. Lawless, and O. Conlan. Dynamic personalisation for digital cultural heritage collections. In *UMAP Workshops – Patch 2013*, 2013.

[5] I. Hickson. Web Storage. W3C Recommendation, 2013. `http://www.w3.org/TR/webstorage/`.

[6] N. Mehta, J. Sicking, E. Graff, A. Popescu, J. Orlow, and J. Bell. Indexed Database API. W3C Candidate Recommendation, 2013. `http://www.w3.org/TR/IndexedDB/`.

[7] G. Oskam, J. Andersson, and Álex Hinojo. Case study: Europeana edit-a-thon. online, 2013.

[8] A. Popescu. Geolocation API Specification. W3C Recommendation, 2013. `http://www.w3.org/TR/geolocation-API/`.

[9] B. J. Rhodes. *Just-In-Time Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[10] T. Ruotsalo, K. Haav, A. Stoyanov, S. Roche, E. Fani, R. Deliai, E. Mäkelä, T. Kauppinen, and E. Hyvönen. Smartmuseum: A mobile recommender system for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 20(0):50 – 67, 2013.

[11] R. Sanderson, P. Ciccarese, and H. V. de Sompel. Open Annotation Data Model. W3C Community Draft, 2013. `http://www.openannotation.org/spec/core/`.

[12] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 718–723, New York, NY, USA, 2006. ACM.